9th DOE Meeting

January 29 2002

Goals, Design & Implementation of a Versatile Microarray Database

Marc Rejali

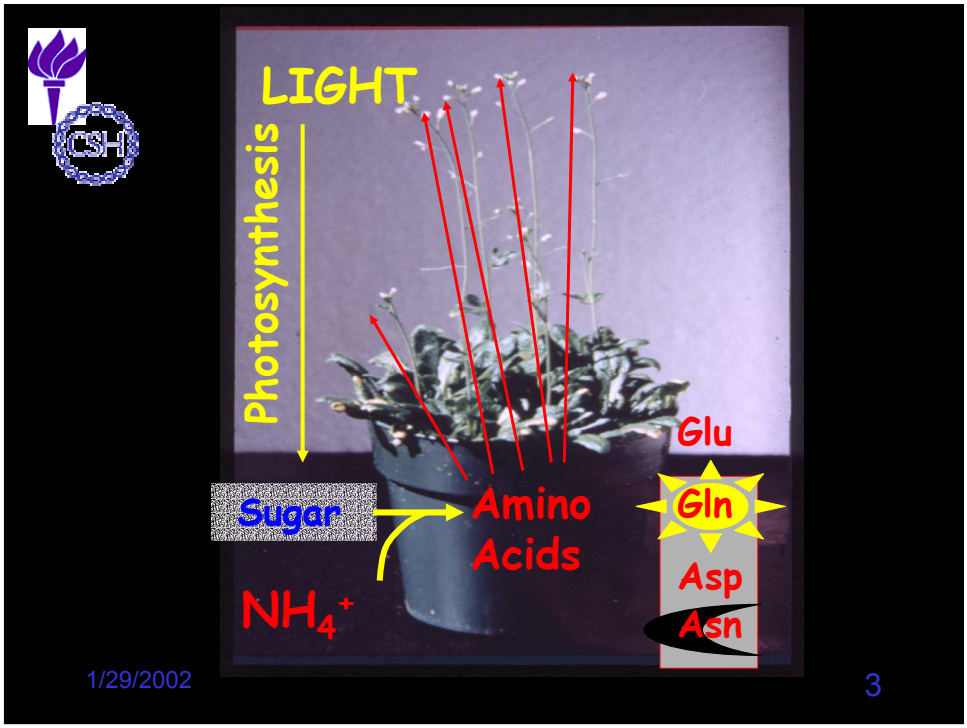
(jointly with M. Antoniotti, V. Cherepinsky, C. Leventhal, S. Paxia, A. Rudra, J. West & B. Mishra)

Department of Computer Science & Mathematics (Courant, NYU)

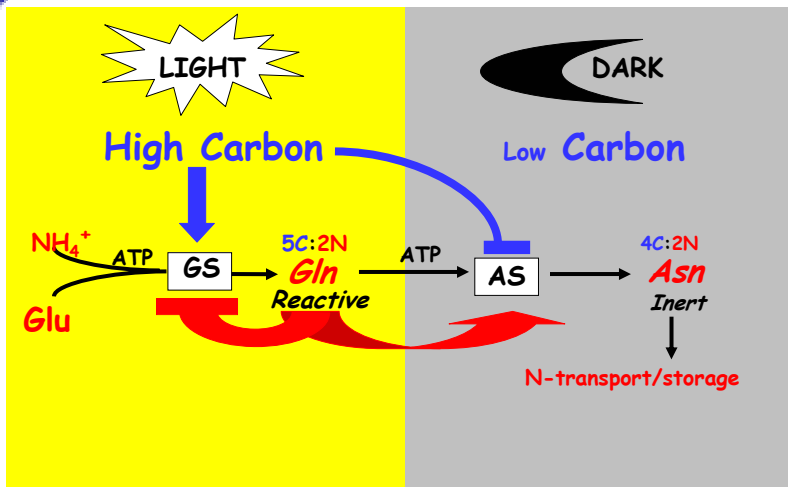
Cold Spring Harbor Laboratory

Various Projects at NYU Based on Microarrays

- Nitrogen Pathway analysis in Arabidopsis
  - (in cooperation with NYU Biology Dept., USC & Univ. of Illinois)
- Hallucinogen effects on brain functions
  - (in collaboration with Mount Sinai School of Medicine)
- Cancer related cell signaling using different cell lines
- Genome Mapping and Probe Placement:
  - Poster #112



Nitrogen Pathway



Goals

- Facilitate collaborative research
  - Data sharing
  - Rich graphical and web interface
- Controlled data visibility/access
  - Group-based read/write access
- General access to published data
  - Simple data retrieval
- Provision of analysis and clustering algorithms
  - Standard methods
  - New methods / enhancements

NYU Microarray Database (NYUMAD).

- The underlying DB schema design
  - follows the specifications put forth by the Microarray Gene Expression Database group (http://www.ebi.ac.uk/microarray/MGED),
  - especially when it comes to the XML-based MAML (Microarray Mark-up Language) exchange format.

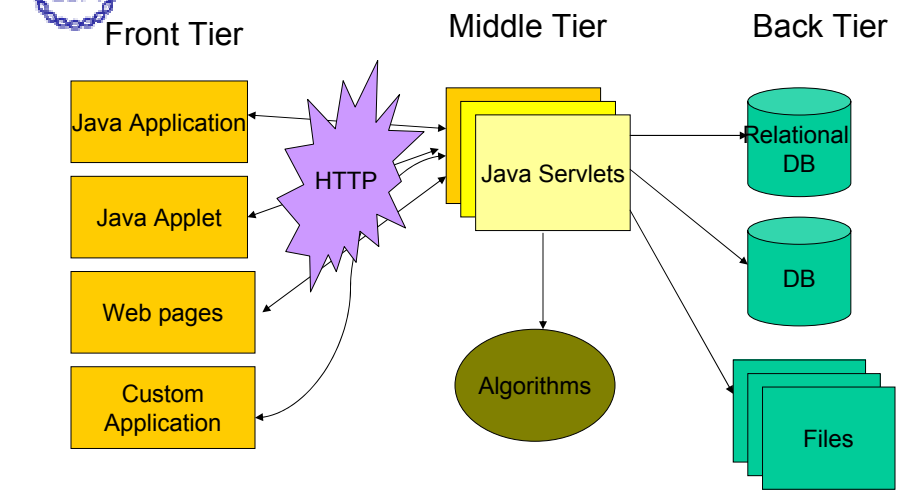
NYUMAD Functionality

- Stores array data, experimental conditions and protocols
- Data is served to "clients" via the world wide web (WWW).
  - Clients can be the NYUMAD Java applet/application
  - custom-built user programs,
  - XML files retrieved using a simple HTTP text based request format.
- The NYUMAD applet presents data in a logical manner and allows easy navigation through the data.
  - Allows straight-forward updating of existing data and the insertion of new data.
  - Retrieval of data in text or XML format

Architecture

- The NYUMAD has a three-tier architecture:
- *Front tier*
  - NYUMAD java applet/application
  - User's custom-built programs
  - HTML forms.
- *Middle tier*
  - Java servlets handle requests and submissions from the front tier
  - Invisible to the end user.
  - Algorithms and analytics (C++, C, other languages)
- *Back tier*
  - Relational database management systems (RDBMS, currently PostgreSQL running on a 6 nodes Linux cluster).
  - File management system used to store large files such as image files.

Architecture



File Edit Data

MAD
- Experiments
- Protocols
- Array Platforms
- Clustering

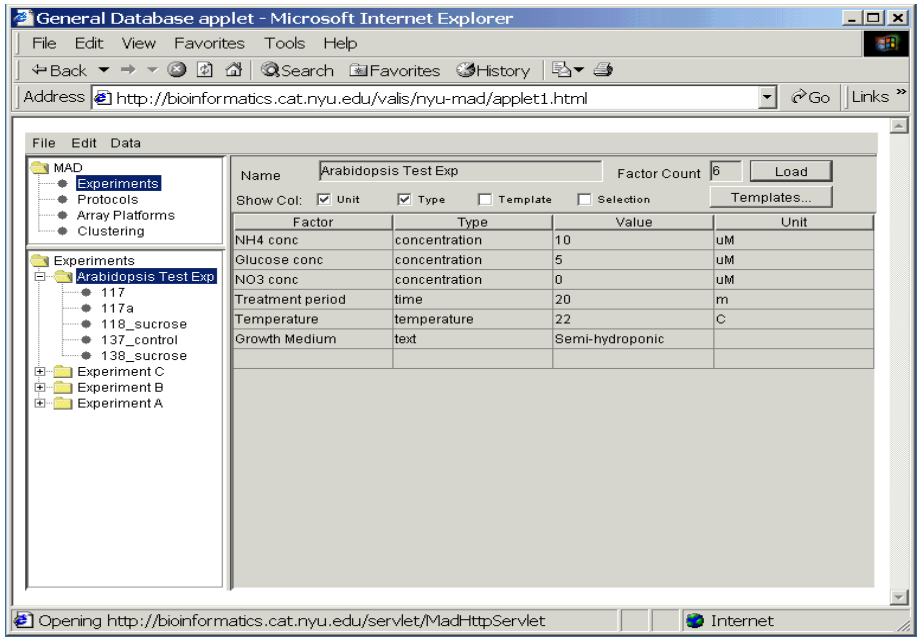
Test Array A1
Affy_1000
Affymetrix Arabidopsis Array
Incyte Arabidopsis chip

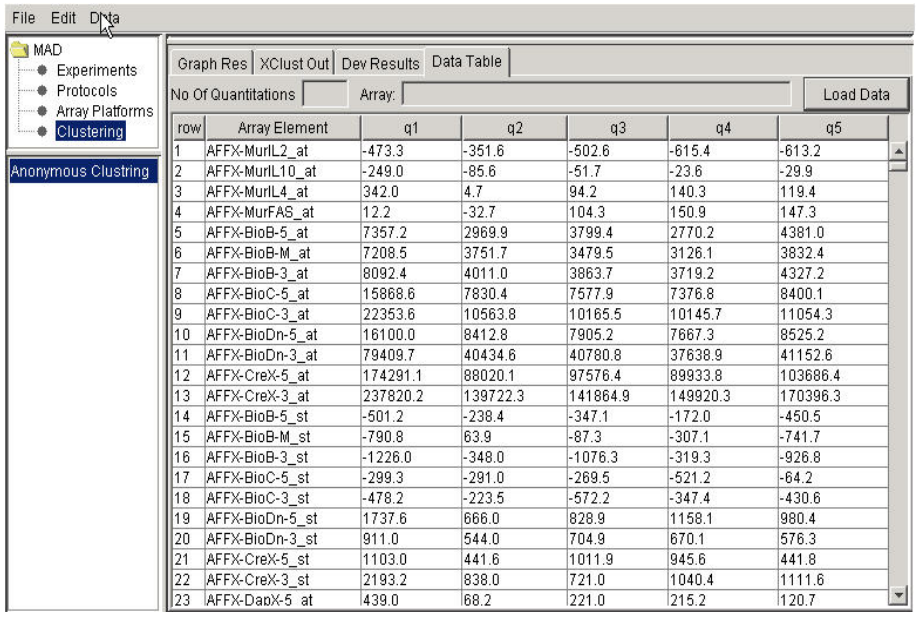
Name: Affymetrix Arabidopsis Array Surface Type:

View | Import

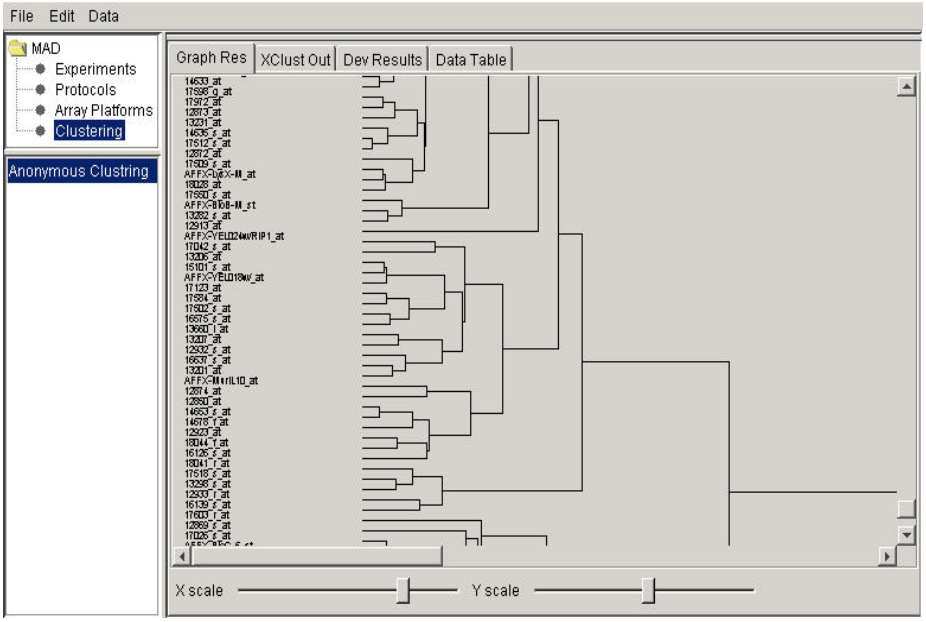
No Of Elements: 8297 Load Elements

Element ID	Description	X	Y
12863_r_at	X53579 A.thaliana agamous (AG) gene.	373	4
12864_s_at	X53579 A.thaliana agamous (AG) gene.	389	3
12869_s_at	"U20184 Arabidopsis thaliana MADS-box protein AGL14 (AGL14) mRNA, part...	469	4
12870_s_at	"U22528 Arabidopsis thaliana AGL15 gene, complete cds."	485	3
12883_at	X98925 A.thaliana mRNA for stromal ascorbate peroxidase.	283	5
12886_s_at	X98926 A.thaliana mRNA for thylakoid-bound ascorbate peroxidase.	331	5
12891_at	"M92354 Arabidopsis thaliana ACC synthase (AtACS-6) mRNA, partial cds."	411	5
12892_g_at	"M92354 Arabidopsis thaliana ACC synthase (AtACS-6) mRNA, partial cds."	427	5
12893_at	X82647 A.thaliana mRNA for Class III ADH.	443	5
12894_g_at	X82647 A.thaliana mRNA for Class III ADH.	459	6
12895_at	AJ002597 Arabidopsis thaliana mRNA for membrane-associated salt-induci...	475	6
12896_at	"AC002330 Arabidopsis thaliana BAC T10P11 from chromosome IV, near 15 ...	491	5
12897_at	X83371 A.thaliana mRNA for cyclin delta-3.	523	5
12898_g_at	X83371 A.thaliana mRNA for cyclin delta-3.	7	7
12899_i_at	X57839 A.thaliana CDC2a gene for p34(cdc2)-like cell cycle protein.	23	7
12900_r_at	X57839 A.thaliana CDC2a gene for p34(cdc2)-like cell cycle protein.	37	8
12902_at	X57840 A.thaliana CDC2b gene for p34(cdc2) cell cycle protein.	184	8
12910_s_at	X84230 A.thaliana mRNA for glucose-6-phosphate dehydrogenase (clone E1...	312	7
12912_at	AJ006021 Arabidopsis thaliana mRNA for putative PRL1 associated protein.	344	7
12913_at	Z50851 A.thaliana athb-8 gene.	360	8
12914_s_at	"L24437 Arabidopsis thaliana (COP1) regulatory protein, complete cds."	392	8
12915_at	"L22874 Arabidopsis thaliana COR8 mRNA, complete cds."	408	7





row	Array Element	q1	q2	q3	q4	q5
1	AFFX-MurIL2_at	-473.3	-351.6	-502.6	-615.4	-613.2
2	AFFX-MurIL10_at	-249.0	-85.6	-51.7	-23.6	-29.9
3	AFFX-MurIL4_at	342.0	4.7	94.2	140.3	119.4
4	AFFX-MurFAS_at	12.2	-32.7	104.3	150.9	147.3
5	AFFX-BioB-5_at	7357.2	2969.9	3799.4	2770.2	4381.0
6	AFFX-BioB-M_at	7208.5	3751.7	3479.5	3126.1	3832.4
7	AFFX-BioB-3_at	8092.4	4011.0	3863.7	3719.2	4327.2
8	AFFX-BioC-5_at	15868.6	7830.4	7577.9	7376.8	8400.1
9	AFFX-BioC-3_at	22353.6	10563.8	10165.5	10145.7	11054.3
10	AFFX-BioDn-5_at	16100.0	8412.8	7905.2	7667.3	8525.2
11	AFFX-BioDn-3_at	79409.7	40434.6	40780.8	37638.9	41152.6
12	AFFX-CreX-5_at	174291.1	88020.1	97576.4	89933.8	103686.4
13	AFFX-CreX-3_at	237820.2	139722.3	141864.9	149920.3	170396.3
14	AFFX-BioB-5_st	-501.2	-238.4	-347.1	-172.0	-450.5
15	AFFX-BioB-M_st	-790.8	63.9	-87.3	-307.1	-741.7
16	AFFX-BioB-3_st	-1226.0	-348.0	-1076.3	-319.3	-926.8
17	AFFX-BioC-5_st	-299.3	-291.0	-269.5	-521.2	-64.2
18	AFFX-BioC-3_st	-478.2	-223.5	-572.2	-347.4	-430.6
19	AFFX-BioDn-5_st	1737.6	666.0	828.9	1158.1	980.4
20	AFFX-BioDn-3_st	911.0	544.0	704.9	670.1	576.3
21	AFFX-CreX-5_st	1103.0	441.6	1011.9	945.6	441.8
22	AFFX-CreX-3_st	2193.2	838.0	721.0	1040.4	1111.6
23	AFFX-DapX-5_at	439.0	68.2	221.0	215.2	120.7



Conclusion

- Functionality ranges from the storage of the data to front-end capabilities for the presentation and maintenance of the data.
- The database is a unified platform to understand the microarray based gene expression data.
  - *The data can be output to a wide class of clustering algorithms, based on various "similarity measures" and various approaches to grouping*
  - *We have developed a new statistically-robust similarity measure based on James-Stein Shrinkage estimators and provided a Bayesian explanation for its superior performance*
  - *Additional research is focused on incorporating statistical tests for validation and measuring the significance (e.g., jackknife and bootstrap tests).*

Further Information

NYU Bioinformatics

http://bioinformatics.cat.nyu.edu

Applet prototype

Requires Java Runtime Environment 1.3

Has guest login – no password required

Email: mcr221@nyu.edu